Creation and Validation of the Spanish Durum Wheat Core Collection

Magdalena Ruiz, Patricia Giraldo, Conxita Royo, and Jose M. Carrillo

ABSTRACT

Spanish wheat (*Triticum* spp.) landraces have a considerable polymorphism, containing many unique alleles, relative to other collections. The existence of a core collection is a favored approach for breeders to efficiently explore novel variation and enhance the use of germplasm. In this study, the Spanish durum wheat (*Triticum turgidum* L.) core collection (CC) was created using a population structure–based method, grouping accessions by subspecies and allocating the number of genotypes among populations according to the diversity of simple sequence repeat (SSR) markers. The CC of 94 genotypes was established, which accounted for 17% of the accessions in the entire collection. An alternative core collection (CH), with the same number of genotypes per subspecies and maximizing the coverage of SSR alleles, was assembled with the Core Hunter software. The quality of both core collections was compared with a random core collection and evaluated using geographic, agromorphological, and molecular marker data not previously used in the selection of genotypes. Both core collections had a high genetic representativeness, which validated their sampling strategies. Geographic and agromorphological variation, phenotypic correlations, and gliadin alleles of the original collection were more accurately depicted by the CC. Diversity arrays technology (DArT) markers revealed that the CC included genotypes less similar than the CH. Although more SSR alleles were retained by the CH (94%) than by the CC (91%), the results showed that the CC was better than CH for breeding purposes.

Abbreviations: A-NE distance, average distance with the nearest genotype; CC, core collection generated with the stratified method; CH, core collection generated with the Core Hunter program; CR, coincidence rate of range; DArT, diversity arrays technology; E-NE distance, distance between each entry in the core collection and the nearest neighboring entry; MD, mean difference percentage; RC, core collection generated with the random sampling method; SC, primary subset collection; SSR, simple sequence repeat; VD, variance difference percentage; VR, variable rate of coefficient of variation.

The concept of core collection was introduced with the aim of increasing the efficiency of characterization and use of the germplasm stored in gene banks, while preserving as much as possible the genetic diversity of the collections (Frankel and Brown, 1984; Brown, 1989a). To achieve this objective, a core collection should maximize the genetic variation contained in the whole collection with a minimum of repetitiveness. Brown (1989a) suggested three genetic criteria to be meet by a good core collection. First, the major subspecific taxa and geographic regions should be included in the core subset. Second, emphasis should be given to broadly adapted rather than intensely specialized alleles. Finally, within the above criteria, genetic diversity, especially as

determined by the number of alleles per locus, should be maximized in the core collection. When creating a core collection, curators should make decisions on: (i) the optimal number of accessions needed to retain an acceptable proportion of alleles present in a given collection, and (ii) the method used to select accessions for the core subset. A number of studies have recommended sizes of the core collection ranging from 5 to 30% of the entire collection to retain a great part of the genetic variability with a manageable number of accessions (Brown, 1989b; Yonezawa et al., 1995; Charmet and Balfourier, 1995; Noirot et al., 1996; Bisht et al., 1998; van Hintum et al., 2000).

Regarding sampling methods, the most common approaches are the M (maximization) strategy (Schoen and Brown, 1993), and stratified sampling (Brown, 1989a; Erskine and Muehlbauer, 1991; Schoen and Brown, 1993). The objective of the M-strategy is to maximize the total allelic diversity in the core collection (as favored by taxonomists and geneticists), whereas the objective of stratified sampling is to include widely adapted accessions that optimize the representativeness of the genetic diversity in the core collection (which is the breeder's preference). In the M-strategy, accessions are directly selected from the whole collection by maximizing the number of observed alleles at each marker locus. In the stratified strategy, the entire collection is divided into groups, which are as genetically distinct as possible, to maximize the diversity among clusters while minimizing the diversity within groups. This strategy uses three steps: (i) grouping accessions using the available information (e.g., passport data); (ii) allocating the number of accessions among groups according to the linear or logarithmic proportion of the group size in the basic collection (Brown, 1989a), or according to the proportion of the diversity in each group (such as the H strategy in Schoen and Brown [1993], the D method in Franco et al. [2006], and the G method in Li et al. [2002]); and (iii) choosing the individuals from each group randomly or according to genetic membership among accessions (such as the clustering method in Li et al. [2002]). Different researchers have recommended the stratified strategy, especially if the user is interested in optimizing the chance of finding material to be used in a breeding program (Spagnoletti and Qualset, 1993; Odong et al., 2011, 2013; Zhang et al., 2011). This method, however, requires a previous knowledge of the genetic differentiation of the collection.

Thachuk et al. (2009) proposed another approach that can be applied with or without stratification of the base collection (Core Hunter program) for selecting core subsets that optimize a single or multiple genetic measures simultaneously. Some studies have shown that Core Hunter can lead to core subsets with increased genetic diversity and improved average genetic distance compared with other methods as the M-strategy (Thachuk et al., 2009; Díez et al., 2012). Core Hunter has additional advantages: repeating the selection process produces a consistent solution, it is freely available, and less time-consuming than other algorithms (Thachuk et al., 2009; Díez et al., 2012).

Core subsets can be formed on the basis of morphological, phenotypic, or molecular marker data. Diverse studies have shown that core collections based on genotypic data—such as single sequence repeats (SSRs)—retain larger genetic variability and have superior representatives than those based only on phenotypic values (Hu et al., 2000; Balfourier et al., 2007). Due to the cumbersome work involved in the analysis of molecular markers for the whole collection, molecular characterization of primary subsets of the collections can help in the creation of core collections. On the other hand, a correct evaluation of the collection quality should be based, when possible, on data not previously used for the selection of the core, such as relevant phenotypic traits or a different type of molecular markers (van Hintum et al., 2000; Parra-Quijano et al., 2011; Odong et al., 2013).

Spanish durum wheat (*Triticum turgidum* L.) landraces have a considerable polymorphism, containing many unique alleles, relative to other germplasm collections (Pflüger et al., 2001; Moragues et al., 2006; Aguiriano et al., 2006, 2008; Ruiz et al., 2012a; Nazco et al., 2013). Thus, the creation of a core collection could be an important tool for breeders to facilitate the intensive study, evaluation, and use of germplasm. In a previous research (Ruiz et al., 2012a), the genetic structure of a primary subset of 190 genotypes representative of the entire collection was assessed. The results revealed that the collection was structured in nine populations with a higher influence on their differentiation of the subspecies taxa with respect to the agroecological zone of origin.

The aims of the present study were: (i) to create the Spanish core collection of *Triticum turgidum* L. by using the stratified method, (ii) to compare the results with those obtained by generating the core collection with the Core Hunter program, and (iii) to evaluate their representativeness.

MATERIALS AND METHODS

Plant Material

The research material used was a primary subset collection (SC) comprising 190 genotypes of three subspecies: 13 of *dicoccon* (Schrank) Thell., 38 of *turgidum*, and 139 of *durum* (Desf.) Husn. (Supplementary Table S1), representatives of the entire collection of 555 Spanish landraces and old cultivars of *Triticum turgidum* L. maintained at the National Plant Genetic Resources Center (CRF-INIA). In a previous study (Ruiz et al., 2012a), the genetic structure of the SC was investigated with SSR markers using the Bayesian clustering implemented in STRUCTURE v 2.1 (Pritchard et al., 2000; Falush et al., 2003). In the same study (Ruiz et al., 2012a), the SC diversity was also characterized using gliadin and diversity arrays

technology (DArT) markers, and for agromorphological traits. The geographic origin of each landrace was recorded and nine agroecological zones of origin were identified.

Creation of the Core Collections

Each accession was represented by a unique genotype derived from a single selected plant, representative of each original accession. A predetermined sampling intensity of 10 to 20% was considered optimal to capture the 70% of the alleles (Brown, 1989b) and most of the genetic diversity of quantitative characters of the entire collection (Noirot et al., 1996). The 10 to 20% of the 555 accessions would produce a core collection size between 55 and 110 genotypes. The manageable number of genotypes for operative evaluations was established in a maximum of 100. The final size of 94 genotypes was obtained using the allele frequency-base method (Crossa et al., 1993) to retain a large proportion of SSR alleles that occur at low frequencies (0.05 and 0.03).

The core collection was constructed with the stratified sampling methodology (CC). The stratification of the accessions was based on the genetic structure of the SC proposed by Ruiz et al. (2012a). The first level of grouping was based on subspecies and the second level was based on the populations obtained using SSR markers. The number of genotypes among the subspecies was allocated according to the linear proportion of the group size for *dicoccon*, and to the logarithmic proportion for *turgidum* and *durum* to compensate for the differences in group sizes and genetic diversity in the SC. The number of genotypes per subspecies was maintained fixed in the core collections generated. The size of the sample to be drawn from each population formed using SSR markers was determined according to the proportion of the SSR gene diversity (H_T; Nei, 1973). H_T values in the populations ranged from 0.31 to 0.71. The genotypes from each population were chosen according to their differences in agroecological zones of origin, avoiding redundant genotypes for SSRs and gliadins. An alternative core collection, the CH, was constructed with the search algorithm implemented by the Core Hunter Program (Thachuk et al., 2009) in a complementary manner with the stratified method. Therefore, after determination of the number of genotypes to be sampled from each taxonomic group (the same as in the CC), Core Hunter was run independently for each subspecies. The Core Hunter algorithm employs an advanced stochastic local search algorithm—replica exchange Monte Carlo—to maximize a pseudo-index that may integrate diversity indexes and genetic distances (Thachuk et al., 2009). In our case, the CH maximized the coverage of SSR alleles present in the SC. Five replicates with a maximum runtime of 100 s were performed for each subset. The profiles of the 190 genotypes in the SC are reported (Supplementary Table S1). The 94 genotypes included in of the CC and CH, are indicated. Additional passport data can be obtained from the Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria database (www.inia.es/crf [accessed 23 Aug. 2013]).

A random core collection (RC) was generated to serve as reference. The genotypes were sampled randomly from each subspecies without replacement.

Evaluation of the Core Collections

The quality of the generated core collections was evaluated using geographic, agromorphological, and molecular (gliadin and DArT) marker data. The agromorphological characterization and the molecular analysis of the SC were performed in a previous research (Ruiz et al., 2012a). Tables 1 and 2 show the qualitative and quantitative agromorphological descriptors recorded according to IBPGR (1985). Differences in qualitative traits between the core collections and the SC were checked by chi-square test ($P < 0.05$) for frequency data. For quantitative characters, the mean, variance, range, and coefficient of variation were calculated (Table 2). A homogeneity test (F test) for variances and a t test for means ($P < 0.05$) were used to compare the differences between the core collections and the SC. We calculated the evaluation parameters proposed by Hu et al. (2000): mean difference percentage (MD), variance difference percentage (VD), coincidence rate of range (CR), and variable rate of coefficient of variation (VR). According to the criteria of Hu et al. (2000), a core collection can be considered representative if no more than 20% of the traits have different means (MD $\leq$ 20), and the coincidence rate of range retained by the core collection is no less than 80% (CR $\geq$ 80%). Relationships between quantitative traits were examined using Pearson correlation coefficients. The Dice distance (Dice, 1945) between genotypes was calculated from the DArT data. For each core collection, the distances between each genotype in the SC and the nearest genotype in the core collection (A-NE), and between each entry in the core collection and the nearest neighboring entry (E-NE) were calculated and averaged over all the genotypes according to Odong et al. (2013). The lower the value of A-NE and the higher of E-NE, the better the representativeness of genetic diversity is in the core set (Odong et al., 2013).

Table 1. Relative frequency (%) of the qualitative agromorphological traits in the primary subset collection (SC) and the core collections obtained with the stratified method (CC), Core Hunter algorithm (CH), and random sampling (RC).

Trait	Class	Frequency			
		SC	CC	CH	RC
Growth habit	Prostrate	15.3	20.2	22.3	23.4
	Intermediate	72.6	66.0	63.8	66.0
	Erect	12.1	13.8	13.8	10.6
Awn barbs	Rough	88.9	93.6	93.6	92.6
	Smooth	11.1	6.4	6.4	7.4
Awnedness	Awnless	0.5	1.1	1.1	0.0
	1–3 cm	9.5	12.8	13.8	16.0
	3–8 cm	4.7	7.4	6.4	6.4
	>8 cm	85.3	78.7	78.7	77.7
Spike density	Lax	3.7	5.3	5.3	4.3
	Intermediate	40.5	50.0	54.3	50.0
	Dense	49.5	36.2	34.0	38.3
	Very dense	6.3	8.5	6.4	7.4
Glume hairiness	Hairless	61.6	64.9	66.0	60.6
	Low	8.9	8.5	9.6	10.6
	High	29.5	26.6	24.5	28.7
Glume color	White	48.9	50.0	48.9	51.1
	Red to brown	35.3	37.2	36.2	29.8
	Black	15.8	12.8	14.9	19.1
Seed color	White	66.8	53.2*	53.2*	60.6
	Red	33.2	46.8*	46.8*	39.4

* Significantly different from SC at the 0.05 probability level.

Table 2. Summary statistics of the quantitative agromorphological traits in the primary subset collection (SC) and in the core collections obtained with the stratified method (CC), Core Hunter algorithm (CH), and random sampling method (RC).

Trait	Mean				Variance				Coefficient of variation				Range			
	SC	CC	CH	RC	SC	CC	CH	RC	SC	CC	CH	RC	SC	CC	CH	RC
Days to heading (d)	174.11	175.12	175.34*	175.09	16.22	17.18	16.12	17.91	2.31	2.37	2.29	2.42	167–184	167–184	168–184	168–184
Days to maturity (d)	213.96	214.36	214.62	214.18	12.39	11.70	9.74	7.63**	1.65	1.60	1.45	1.29	201–223	201–223	201–223	205–218
Grain filling period (d)	39.85	39.15	39.28	39.10	15.60	14.94	16.80	17.94	9.91	9.87	10.44	10.83	30–48	30–48	30–48	30–48
Plant height (cm)	124.25	125.55	125.80	125.71	98.22	111.65	118.10	126.94	7.98	8.42	8.64	8.96	95–150	95–150	95–150	97–150
Spike length (cm)	100.10	104.07*	104.77*	102.26	216.37	256.87	242.18	202.34	14.69	15.40	14.85	13.91	72–163	72–163	75–163	75–147
Spikelets/spike (no.)	23.17	23.78	23.74	23.13	12.67	21.72**	21.72**	5.73**	15.36	19.60	19.63	10.35	19–61	19–61	19–61	19–32
100-grain weight (g)	5.78	5.86	5.96	5.96	3.93	5.85*	6.19*	5.77*	34.25	41.30	41.77	40.32	3.5–14.5	3.5–14.5	3.6–14.5	3.59–14.52
Evaluation parameters†	MD	14	29	0	VD	14	14	43	VR	108.28	107.35	98.00	CR	100	99	80

* Significantly different from SC at the 0.05 probability level.

** Significantly different from SC at the 0.01 probability level.

† MD, mean difference percentage; VD, variance difference percentage; VR, variable rate of coefficient of variation; CR, coincidence rate of range.

Table 3. Distribution of the simple sequence repeat alleles in the primary subset collection (SC) and the core collections obtained with the stratified method (CC), Core Hunter algorithm (CH), and random sampling method (RC).

Frequency	No. alleles				Alleles lost		
	SC	CC	CH	RC	CC	CH	RC
>0.1	108	108	108	108	0	0	0
>0.05	186	186	186	186	0	0	0
≥0.03	253	253	253	252	0	0	1
>0.01	395	392	393	383	3	2	12
≤0.01	246	191	210	152	55	36	93
Total	641	583	603	535	58	38	106

RESULTS

Creation of the Core Collections

The analyses of the genetic structure of the primary SC, performed in a previous study (Ruiz et al., 2012a), indicated that the differentiation was greater among subspecies (*dicoccon*, *turgidum*, and *durum*) and among populations defined by SSRs than among the agroecological zones in which the accessions were collected. Hence, we have used the stratified sampling strategy with grouping of genotypes based on taxonomic subspecies and populations obtained using SSR markers. The number of alleles with frequencies higher than 0.03 was 253 (Table 3), which gave a mean of 6 alleles locus^{-1} for the 39 SSR loci analyzed. According to Crossa et al. (1993), the sample size of the core collection established in 94 accessions would be expected to include all these alleles. The final number of genotypes in the CC for each subspecies was 10, 32, and 52 for *dicoccon*, *turgidum*, and *durum*, respectively. The number of genotypes from each population ranged from 2 to 18. A core collection (CH) of the same size and number of genotypes per subspecies was obtained with the Core Hunter software. Core Hunter was run independently on each subspecies with the objective of maximizing the allele coverage in each taxonomic group. Eight accessions of *dicoccon*, 28 of *turgidum*, and 33 of *durum* coincided in both collections. An additional RC collection with the same number of genotypes per subspecies than CC and CH was also generated.

Table 4 shows the number of alleles per SSR locus in the core collections. The genetic parameters H_T and number of effective alleles were better in the core collections than in the SC. The main difference between the CC and the CH was for the allele coverage, which was higher in the CH than in the CC. A more precise analysis of the alleles retained is shown in Table 3. Both core collections captured all the predominant ($P > 0.1$), common ($P \geq 0.05$), and less frequent ($P \geq 0.03$) alleles. The CC included the 78% and the CH 85% of the rare alleles ($P \leq 0.01$), most of them present in only one genotype. The RC performed worse than the CC and the CH in preserving SSR alleles.

Table 4. Gene diversity of each simple sequence repeat locus in the primary subset collection (SC) and in the core collections obtained with the stratified method (CC), Core Hunter algorithm (CH), and random sampling method (RC).

	No. alleles				Effective no. alleles			
Locus	SC	CC	CH	RC	SC	CC	CH	RC
BARC0055	10	9	9	7	3.87	3.87	4.06	4.09
BARC0080	17	15	15	12	8.07	9.00	9.39	8.29
BARC0155	11	7	11	9	3.00	3.31	3.46	3.29
BARC1032	8	7	8	6	1.50	1.76	1.99	1.82
BARC1077	4	3	3	3	2.19	1.86	1.76	1.86
CFA2219	19	16	18	13	8.64	9.58	10.09	8.98
CFA2257	15	14	15	14	2.40	3.47	3.34	3.45
CFA2263	14	12	13	10	3.93	4.07	4.27	3.59
WMC468	5	4	4	5	1.99	1.89	2.02	2.17
WMC522	32	31	30	28	17.58	19.13	21.45	18.49
X0002	4	4	4	4	1.56	1.60	1.66	1.83
X0011	15	14	14	13	8.18	8.53	8.87	8.29
X0018	12	12	12	10	3.63	5.26	5.61	4.70
X0046	15	15	15	13	8.33	9.63	10.05	9.39
X0060	19	16	17	14	4.98	7.00	6.64	6.21
X0088	17	15	16	15	9.86	8.33	8.43	8.82
X0095	6	6	6	5	3.53	3.56	3.66	3.23
X0099	11	10	11	10	5.10	4.85	4.79	4.68
X0136	41	34	35	32	18.11	19.50	26.64	19.36
X0148	12	12	12	9	4.59	5.16	4.99	4.70
X0154	16	13	13	13	5.01	5.93	5.58	5.44
X0155	16	14	15	10	6.88	5.63	5.94	5.13
X0156	20	18	18	19	7.38	7.48	7.46	8.12
X0181	22	21	21	20	11.22	14.58	13.22	11.56
X0186	28	25	28	25	5.07	6.57	8.06	6.51
X0234	19	19	19	12	4.35	6.21	5.51	4.35
X0251	11	11	11	11	3.95	4.19	4.49	3.75
X0299	21	20	20	18	7.05	9.14	8.84	7.09
X0312	26	26	26	22	4.43	9.15	9.56	6.83
X0332	20	18	19	19	5.90	7.39	7.44	6.39
X0389	14	13	14	13	6.69	6.93	7.45	7.39
X0408	17	17	17	13	3.96	4.76	4.76	3.86
X0459	23	20	22	18	7.23	9.02	10.27	7.88
X0494	9	8	9	9	4.06	4.60	4.62	5.01
X0513	4	4	4	4	2.45	2.90	2.59	2.87
X0570	15	14	15	13	5.20	5.87	6.58	5.15
X0577	41	37	35	37	18.94	22.69	21.47	20.45
X0601	14	12	11	12	4.70	5.87	5.90	5.00
X0604	18	17	18	15	5.22	6.42	6.92	5.63
Mean	16.44	14.95	15.46	13.72	6.17	7.09	7.75	6.55
Total	641	583	603	535	240.71	276.66	289.83	255.63
Allele coverage	100	0.91	0.94	0.84				
H_T†	0.77	0.79	0.80	0.78				

†Total gene diversity.

Validation of the Core Collections

Geographical ranges for latitude (28° N to 43°29′39″ N) and longitude (16° W to 2°6′19″ E) were similar in the SC, CC, and CH. For RC, the latitude was in the range of 28° N to 43°27′57″ N. The altitude ranged from 6 to 1065 m in the SC, from 6 to 1033 m in the CC, from 6 to 1020 m in the CH, and from 10 to 1065 m in the RC. All nine agroecological zones were included in the CC and RC. For the CH, one agroecological zone (northwest of Spain) was not represented for the subspecies *durum*.

Frequency distributions of the qualitative agromorphological traits were not affected by selection, except for seed color in the CC and CH (Table 1). The RC failed to capture awnless spikes. For the quantitative traits, significant differences among means were recorded only for spike length in the CC, and for spike length and days to heading in the CH (Table 2). Variances for number of spikelets per spike and 100-grain weight were significantly higher in the core collections than in the SC, indicating that they captured more variation. The range for all the traits was retained in the CC (Table 2), but those for days to heading, spike length, and 100-grain weight were not completely included in the CH. The coefficients of variation were similar or higher in the CC and CH than in the SC, except for days to maturity in the CC, and days to heading and maturity in the CH. The evaluation parameter values of MD, VR, and CR were better in the CC than in CH (Table 2). The RC showed the worst values for all the evaluation parameters, except for MD. Lower variance, CV, and range of variation were found in the RC for some traits. All the high ($r > 0.40$) and significant ($P < 0.05$) correlations observed in the SC were conserved in the CC, except for that between plant height and days to heading in the subspecies *dicoccon*. The CH failed to capture the correlations between grain filling period and days to maturity, plant height and days to heading, and spike length and days to heading in *dicoccon*, and between days to heading and maturity in *durum*. The RC also failed to retain most of the relevant correlations found in the three subspecies.

All the gliadin alleles were present in the CC and CH, except for two and three alleles, respectively, at the *Gli-2* loci (Table 5). The number of effective alleles and the H_T values were higher in the CC and CH than in the SC. The RC showed worse values than CC and CH for the genetic indexes evaluated. The DArT data were analyzed separately for each species and the four collections. The A-NE and E-NE distances were, respectively, 0.11 and 0.44 in the CC, 0.12 and 0.39 in the CH, and 0.21 and 0.32 in the RC. The minimum distance between genotypes was 0.00 in the SC and RC, and 0.21 and 0.05 in the CC and CH, respectively. The CC also showed the largest value for the minimum distance in each of the three subspecies.

DISCUSSION

The primary goal of this study was the development of the Spanish core collection of *Triticum turgidum* L. as a useful tool for plant breeders. This collection should represent the genetic and phenotypic variability included in the entire

Table 5. Gene diversity of each gliadin locus in the primary subset collection (SC) and in the core collections obtained with the stratified method (CC), Core Hunter algorithm (CH), and random sampling method (RC).

	No. alleles				Effective no. alleles			
Locus	SC	CC	CH	RC	SC	CC	CH	RC
Gli-A1	11	11	11	10	4.12	4.96	4.90	4.55
Gli-B1	12	12	12	12	3.33	4.24	4.34	4.05
Gli-A2	17	17	16	17	8.84	7.38	6.81	7.34
Gli-B2	27	25	25	21	7.57	11.97	12.70	9.75
Mean	16.75	16.25	16.00	15.00	5.96	7.14	7.19	6.42
Total	67	65	64	60	23.85	28.55	28.75	25.68
Allele coverage	100	0.97	0.96	0.90				
H_T[†]	0.80	0.84	0.84	0.82				

[†]Total gene diversity.

collection. For this purpose, we selected a primary subset collection of 190 accessions characterized using 39 highly variable, discriminating, and broadly distributed SSR loci that yielded a total of 641 alleles (Ruiz et al., 2012a). Our strategy used a population structure–based method, grouping accessions by subspecies and SSR populations (Ruiz et al., 2012a) in conjunction with the diversity index H_T within populations. This method seems to increase the retention of diverse genotypes with significant allele richness, which combines the objectives of both geneticists and breeders (Franco et al., 2006; Zhang et al., 2011).

The high genetic variability of this collection is notable (Ruiz et al., 2012a), which highlights the complexity of designing a core collection of a suitable size to capture the SSR diversity present in the entire collection. Zhang et al. (2011) proposed two principles that should be considered in determining the size of a core collection. The first relates to sampling efficiency and aims to find a "point of compromise" between gain of genetic variation and elimination of genetic redundancy in the core collection. The second relates to the sampling validity and aims to find a "point" at which the prevalent variation types can be preserved into the core collection. In the present study, we tried to balance both principles fixing the first one and evaluating the quality of the collection with phenotypic and other different genotypic data than the SSR markers used in the selection of the core. The size of the core collection of 94 genotypes represented a 17% of the entire collection (555 accessions), and it is within the range from 5 to 20% proposed by different authors to retain at least the 70% of the variability (Brown, 1989a). The results revealed that our CC preserved the 91% of the SSR alleles observed in the SC. This value is in agreement with other studies that have reported levels between 70 and 97% of SSR allelic retention in core collections (Balfourier et al., 2007; Zhang et al., 2011; Odong et al., 2011; Jewell et al., 2012). All the alleles with $P \geq 0.03$ and the 99% of alleles with $P > 0.01$ were captured (Table 3). This implies that the CC included all the predominant and common alleles, but there were differences in rare alleles, most of them represented by only one accession. Localized, low-frequency alleles are of little interest in genetic conservation; they would not contribute to the genetic diversity needed to develop elite cultivars and, therefore, their inclusion in the CC may not be worthwhile (Marshall and Brown, 1975; Frankel et al., 1995; Odong et al., 2011; Zhang et al., 2011).

In addition to the comparison with the whole collection, the evaluation of a core collection usually involves contrasting cores created using different methods. We created an alternative core collection of 94 genotypes with the software Core Hunter and a random sampling core collection. Core Hunter could lead to core subsets that accurately represent both SSR diversity and morphological variability (Díez et al., 2012). The CH had the same number of genotypes per subspecies than the CC, but the allele coverage was maximized in each subspecies, which allowed retaining the 94% of the alleles. Both core collections shared the 73% of the accessions (Supplementary Table S1).

To evaluate the representativeness of the CC, the geographic, phenotypic, and genotypic (gliadins and DArTs) variability included in the collections was compared. These evaluation data were not used in the sampling of the cores following the recommendations by Van Hintum el al. (2000), Parra-Quijano et al. (2011), and Odong et al. (2013). The results showed that the CC and CH had a high phenotypic and genetic representativeness much higher than the RC, which validated both sampling strategies. The high values of VD and VR for agromorphological traits and those of H_T for molecular markers (Tables 2, 4, and 5) showed that the CC and CH had less redundancy than the SC. However, some small differences were found between the two core collections. In this sense, the CC retained slightly better geographic and agromorphological diversity compared to the CH (Tables 1 and 2). The mean values of MD < 20 and CR > 80 in the CC compared with the MD > 20 in the CH indicated that phenotypic variation of the original collection was better accurately depicted by the CC (Hu et al., 2000). More of the phenotypic correlations found in the SC were maintained in the CC than in the CH, which showed that most of the coadapted gene complexes governing these traits were properly sampled and preserved in the CC. All these results revealed that the CC represented the patterns of phenotypic variation of the SC better than the CH.

Gliadin and DArTs are valuable molecular markers for the quantification of genetic diversity in wheat (Metakovsky et al., 1991; Aguiriano et al., 2006; Stodart et al., 2007; White et al., 2008; Raman et al., 2010; Dreisigacker et al., 2012). The DArT markers offer a deep genome coverage and the power to detect even very low level of polymorphisms. Three distances were calculated with the DArT markers: the A-NE, E-NE, and the minimum distance

between genotypes. The A-NE distance indicates how well each accession in the whole collection is represented in the core collection. It seems to be a very good criterion to evaluate core collections formed for the purposes of maximizing the representativeness of genetic diversity (Odong et al., 2013). The low values obtained in the CC and CH showed that they maximized the representativeness of the genetic diversity of the SC. The E-NE distance is a good criterion to evaluate that the core collection has entries that are as different as possible from each other (Odong et al., 2013). The higher values detected in the CC than in CH indicated that the CC represented a larger range of genotypes than CH. Based on the minimum distances between genotypes, the CC included genotypes less similar than those included in CH which fits better with breeders' objectives. The CC also captured more gliadin alleles than CH (Table 5). The evaluation of the RC with gliadin and DArT markers confirmed that this collection had less quality than the CC and CH.

Comparisons of the SSR alleles retained in both core collections indicated that the CH captured more SSR rare alleles (Table 3) and performed better concerning the taxonomists' interest. In fact, Core Hunter algorithm could also select a different core collection of 94 genotypes with optimal allele coverage of 97%. This collection, with four more *durum* genotypes instead of four of *turgidum*, retained fewer gliadin alleles (four lost alleles) and lower phenotypic variability (for days to heading and to maturity, spike length, and 100-grain weight).These results are in agreement with the fact that the Core Hunter currently considers only genetic data.

The main purpose for defining core subsets is to ensure that plant germplasm collections will be used in such a way that they provide efficient access to the entire range of genetic variation. This would facilitate the efficiency of preliminary germplasm evaluations for traits of interest (Crossa et al., 1993). In a recent study (Ruiz et al., 2012b) 51 accessions from CC and CH were evaluated for mineral (iron [Fe] and zinc [Zn]) content in whole grain. The results showed that the genotypes of the CC possessed higher variation ranges than those obtained in 13 modern cultivars, and similar to those obtained in 102 Spanish landraces. The evaluated genotypes of the CH had the same range of variation for Zn as genotypes than those from the CC, but the range of variation was lower for Fe. The analysis evidenced the wide variability of the CC, as well as the potential of these genotypes to be incorporated into breeding programs.

Core Hunter algorithm has demonstrated to be a fast and powerful method for designating core subsets, especially in absence of knowledge of a clear genetic structure in the base collection. However, selection of crop varieties always depends on phenotypic traits and a sole focus on genetic information may bias results due to nonfunctional genetic variations (Thachuk et al., 2009). The results presented here showed that the core collection designed with the stratified method in conjunction with a diversity index well combines the representation of genetic and phenotypic variability. This CC includes a broad range of adapted genotypes, maximizing the representativeness of the genetic diversity in the initial collection, and could be extremely useful for a more efficient use of genetic resources in breeding.

Acknowledgments

This research was supported by the project RF2006-00020-C03 from the National Institute for Agricultural and Food Research and Technology of Spain, the project AGL-2012-38345 from the Ministry of Economy and Competitiveness, and the European Fund for Regional Development (FEDER). We thank Dr. C.M. Diez and Dr. C. Thachuk for their help with the Core Hunter program. The Centre University of Lleida–Institute for Food and Agricultural Research and Technology forms part of the Centre CONSOLIDER INGENIO 2010 on Agrigenomics funded by the Spanish Ministry of Education and Science.

References

Aguiriano, E., M. Ruiz, R. Fite, and J.M. Carrillo. 2006. Analysis of genetic variability in a sample of the durum wheat (*Triticum durum* Desf.) Spanish collection based on gliadin markers. Genet. Resour. Crop Evol. 53:1543–1552. doi:10.1007/s10722-005-7767-z

Aguiriano, E., M. Ruiz, R. Fite, and J.M. Carrillo. 2008. Genetic variation for glutenin and gliadins associated with quality in durum wheat (*Triticum turgidum* L. ssp. *turgidum*) landraces from Spain. Span. J. Agric. Res. 6:599–609.

Balfourier, F., V. Roussel, P. Strelchenko, F. Exbrayat-Vinson, P. Sourdille, G. Boutet, J. Koenig, C. Ravel, O. Mitrofanova, M. Beckert, and G. Charmet. 2007. A worldwide bread wheat core collection arrayed in a 384-well plate. Theor. Appl. Genet. 114:1265–1275. doi:10.1007/s00122-007-0517-1

Bisht, I.S., R.K. Mahajan, T.R. Lokknathan, and R.C. Agrawal. 1998. Diversity in Indian sesame collection and stratification of germplasm accessions in different diversity groups. Genet. Resour. Crop Evol. 45:325–335. doi:10.1023/A:1008652420477

Brown, A.H.D. 1989a. Core collections: A practical approach to genetic resources management. Genome 31:818–824. doi:10.1139/g89-144

Brown, A.H.D. 1989b. The case for core collections. In: A.H.D. Brown et al., editors, The use of plant genetic resources. Cambridge Univ. Press, Cambridge, UK. p. 136–156.

Charmet, G., and F. Balfourier. 1995. The use of geostatistics for sampling a core collection of perennial ryegrass populations. Genet. Resour. Crop Evol. 42:303–309. doi:10.1007/BF02432134

Crossa, J., C.M. Hernandez, P. Bretting, S.A. Eberhart, and S. Taba. 1993. Statistical genetic considerations for maintaining germplasm collections. Theor. Appl. Genet. 86:673–678. doi:10.1007/BF00222655

Dice, L.R. 1945. Measures of the amount of ecologic association between species. Ecology 26:297–302. doi:10.2307/1932409

Díez, C.M., A. Imperato, L. Rallo, D. Barranco, and I. Trujillo. 2012. Worldwide core collection of olive cultivars based on simple sequence repeat and morphological markers. Crop Sci. 52:211–221. doi:10.2135/cropsci2011.02.0110

Dreisigacker, S., H. Shewayrga, C. Crossa, V.N. Arief, I.H. DeLacy, R.P. Singh, M.J. Dieters, and H.-J. Braun. 2012. Genetic structures of the CIMMYT international yield trial targeted to irrigated environments. Mol. Breed. 29:529–541. doi:10.1007/s11032-011-9569-7

Erskine, W., and F.J. Muehlbauer. 1991. Allozyme and morphological variability, outcrossing rate and core collection formation in lentil germplasm. Theor. Appl. Genet. 83:119–125. doi:10.1007/BF00229234

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Franco, J., J. Crossa, M.L. Warburton, and S. Taba. 2006. Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. Crop Sci. 46:854–864. doi:10.2135/cropsci2005.07-0201

Frankel, O.H., and A.H.D. Brown. 1984. Plant genetic resources today: A critical appraisal. In: J.H.W. Holden and J.T. Williams, editors, Crop genetic resources: Conservation and evaluation. George Allen and Unwin, London. p. 249–257.

Frankel, O.H., A.H.D. Brown, and J.J. Burdon. 1995. The conservation of plant biodiversity. Cambridge Univ. Press, Cambridge, UK.

Hu, J., J. Zhu, and H.M. Xu. 2000. Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. Theor. Appl. Genet. 101:264–268. doi:10.1007/s001220051478

IBPGR. 1985. Revised descriptor list for wheat (*Triticum* ssp). Int. Board for Plant Genet. Resour., Rome.

Jewell, M.C., Y. Zhou, D.S. Loch, I.D. Godwin, and C.J. Lambrides. 2012. Maximizing genetic, morphological, and geographic diversity in a core collection of Australian Bermuda grass. Crop Sci. 52:879–889. doi:10.2135/cropsci2011.09.0497

Li, Z.C., H.L. Zhang, Y.W. Zeng, Z.Y. Yang, S.Q. Shen, C.Q. Sun, and X.K. Wang. 2002. Studies on sampling strategies for establishment of core collection of rice landrace in Yunnan, China. Genet. Resour. Crop Evol. 49:67–74. doi:10.1023/A:1013855216410

Marshall, D.R., and A.H.D. Brown. 1975. Optimum sampling strategies in genetic conservation. In: O.H. Franked and J.G. Hawkes, editors, Crop genetic resources for today and tomorrow. Cambridge Univ. Press, Cambridge, UK. p. 53–80.

Metakovsky, E.V., D. Knezevic, and B. Javornik. 1991. Gliadin allele composition of Yugoslav winter wheat cultivars. Euphytica 54:285–295.

Moragues, M., J. Zarco-Hernandez, M.A. Moralejo, and C. Royo. 2006. Genetic diversity of glutenin protein subunits composition in durum wheat landraces [*Triticum turgidum* ssp. *turgidum* convar. *durum* (Desf.) MacKey] from the Mediterranean basin. Genet. Resour. Crop Evol. 53:993–1002. doi:10.1007/s10722-004-7367-3

Nazco, R., R.J. Peña, K. Ammar, D. Villegas, J. Crossa, M. Moragues, and C. Royo. 2013. Variability in glutenin subunit composition of Mediterranean durum wheat germplasm and its relationship with gluten strength. J. Agric. Sci. doi:10.1017/S0021859613000117 (in press).

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70:3321–3323. doi:10.1073/pnas.70.12.3321

Noirot, M., S. Hamon, and F. Anthony. 1996. The principal component scoring: A new method of constituting a core collection using quantitative data. Genet. Resour. Crop Evol. 43:1–6. doi:10.1007/BF00126934

Odong, T.L., J. Jansen, F.A. van Eeuwijk, and T.J.L. van Hintum. 2013. Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor. Appl. Genet. 126:289–305. doi:10.1007/s00122-012-1971-y

Odong, T.L., J. van Heerwaarden, J. Jansen, T.J.L. van Hintum, and F.A. van Eeuwijk. 2011. Statistical techniques for defining reference sets of accessions and microsatellite markers. Crop Sci. 51:2401–2411. doi:10.2135/cropsci2011.02.0095

Parra-Quijano, M., J.M. Iriondo, E. Torres, and L. De la Rosa. 2011. Evaluation and validation of ecogeographical core collections using phenotypic data. Crop Sci. 51:694–703. doi:10.2135/cropsci2010.05.0273

Pflüger, L.A., L.M. Martin, and J.B. Alvarez. 2001. Variation in the HMW and LMW glutenin subunits from Spanish accessions of emmer wheat (*Triticum turgidum* ssp. *dicoccom* Schrank). Theor. Appl. Genet. 102:767–772. doi:10.1007/s001220051708

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Raman, H., B.J. Stodart, C. Cavanagh, M. Mackay, M. Morell, A. Milgate, and P. Martin. 2010. Molecular diversity and genetic structure of modern and traditional landrace cultivars of wheat (*Triticum aestivum* L.). Crop Pasture Sci. 61:222–229. doi:10.1071/CP09093

Ruiz, M., P. Giraldo, C. Royo, D. Villegas, M.J. Aranzana, and J.M. Carrillo. 2012a. Diversity and genetic structure of a collection of Spanish durum wheat landraces. Crop Sci. 52:1–14. doi:10.2135/cropsci2012.02.0081

Ruiz, M., P. Giraldo, C. Royo, D. Villegas, E. Benavente, M.J. Aranzana, and J.M. Carrillo. 2012b. Genetic structure of a Spanish durum wheat landrace collection. In: Proceedings of the 19th Eucarpia General Congress. Plant Breeding for Future Generations, Budapest. 21–24 May 2012. Hungarian Acad. of Sci., Martonvásár, Hungary. p. 269.

Schoen, D.J., and A.H.D. Brown. 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. USA 90:10623–10627. doi:10.1073/pnas.90.22.10623

Spagnoletti, P.L.Z., and C.O. Qualset. 1993. Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. Theor. Appl. Genet. 87:295–304. doi:10.1007/BF01184915

Stodart, B.J., M.C. Mackay, and H. Raman. 2007. Assessment of molecular diversity in landraces of bread wheat (*Triticum aestivum* L.) held in an ex-situ collection with diversity array technology (DArT TM). Aust. J. Agron. Res. 58:1174–1182. doi:10.1071/AR07010

Thachuk, C., J. Crossa, J. Franco, S. Dreisigacker, M. Warburton, and G.F. Davenport. 2009. Core Hunter: An algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinformatics 10:243. doi:10.1186/1471-2105-10-243

van Hintum, T.J.L., A.H.D. Brown, C. Spillane, and T. Hodgkin. 2000. Core collections of plant genetic resources. IPGRI Tech. Bull. 3. Int. Plant Genet. Resour. Inst., Rome.

White, J., J.R. Law, I. Mackay, K.J. Chalmers, J.S.C. Smith, A. Kilian, and W. Powell. 2008. The genetic diversity of UK, US and Australian cultivars of *Triticum aestivum* measured by DArT markers and considered by genome. Theor. Appl. Genet. 116:439–453. doi:10.1007/s00122-007-0681-3

Yonezawa, K., T. Nomura, and H. Morishima. 1995. Sampling strategies for use in stratified germplasm collections. In: T. Hodgkin et al., editors, Core collections of plant genetic resources. IPGRI, Wiley, West Sussex, UK. p. 35–54.

Zhang, H., D. Zhang, M. Wang, J. Sun, Y. Qi, J. Li, X. Wei, L. Han, Z. Qiu, S. Tang, and Z. Li. 2011. A core collection and mini core collection of *Oryza sativa* L. in China. Theor. Appl. Genet. 122:49–61. doi:10.1007/s00122-010-1421-7